Enabling the Usability of Earth Science Data Products and Services by Evaluating, Describing, and Improving Data Quality throughout the Data Lifecycle

Robert R. Downs[1] , Ge Peng[2], Yaxing Wei[3],
Hampapuram Ramapriyan[4], and David F Moroni[5]

[1] *rdowns@ciesin.columbia.edu*
NASA Socioeconomic Data and Applications Center (SEDAC), CIESIN, Columbia University
[2] ge.peng@noaa.gov
NC State University, Asheville, NC,
[3] weiy@ornl.gov Oak Ridge
National Laboratory, Oak Ridge, TN,
[4] hampapuram.ramapriya@ssaihq.com
Science Systems and Applications, Inc., Lanham, MD,
[5] david.f.moroni@jpl.nasa.gov
NASA Jet Propulsion Laboratory, Pasadena, CA

2015 Fall AGU Meeting
IN14A: Approaches to Improved Collection and Dissemination of Earth Science Data Quality Information I
Monday, 14 December 2015, IN14A-02, 16:15 - 16:30, Moscone West - 2020.

Government support is acknowledged for the authors’ work.
R. R. Downs: NASA contract NNG13HQ04C for the Socioeconomic Data and Applications Distributed Active Archive Center (DAAC);
G. Peng: NOAA Cooperative Agreement NA14NES432003;
Y. Wei: NASA contract NNG14HH39I for the Oak Ridge National Laboratory (ORNL) DAAC;
H. K. Ramapriyan: NASA contract NNG15HQ01C;
D. F. Moroni: NASA contract with the Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA.

Overview

- What is data quality and why do we need data quality assessments?
- Data quality in standards and guidelines
- Data quality in data lifecycle and workflows
- Identified actions and implementations for improving data quality
- Implications and progressive planning for data quality assessments
- Recommendations and taking action

What is data quality and why is it important?

- Assessment of the potential usefulness of data and metadata
  - Data quality can be assessed for multiple purposes
  - Metadata often includes information on data quality characteristics to enhance overall data quality through self-description
- Data quality offers value to enable use
  - Quality attained for internal use might be insufficient for external use
  - Enabling public use of data may require higher levels of data quality
  - Investments in data curation become worthwhile when data are used
- Potential users need to determine potential usefulness of data
  - An assessment of data quality is necessary to determine usefulness
  - Each potential use of data could require a different assessment

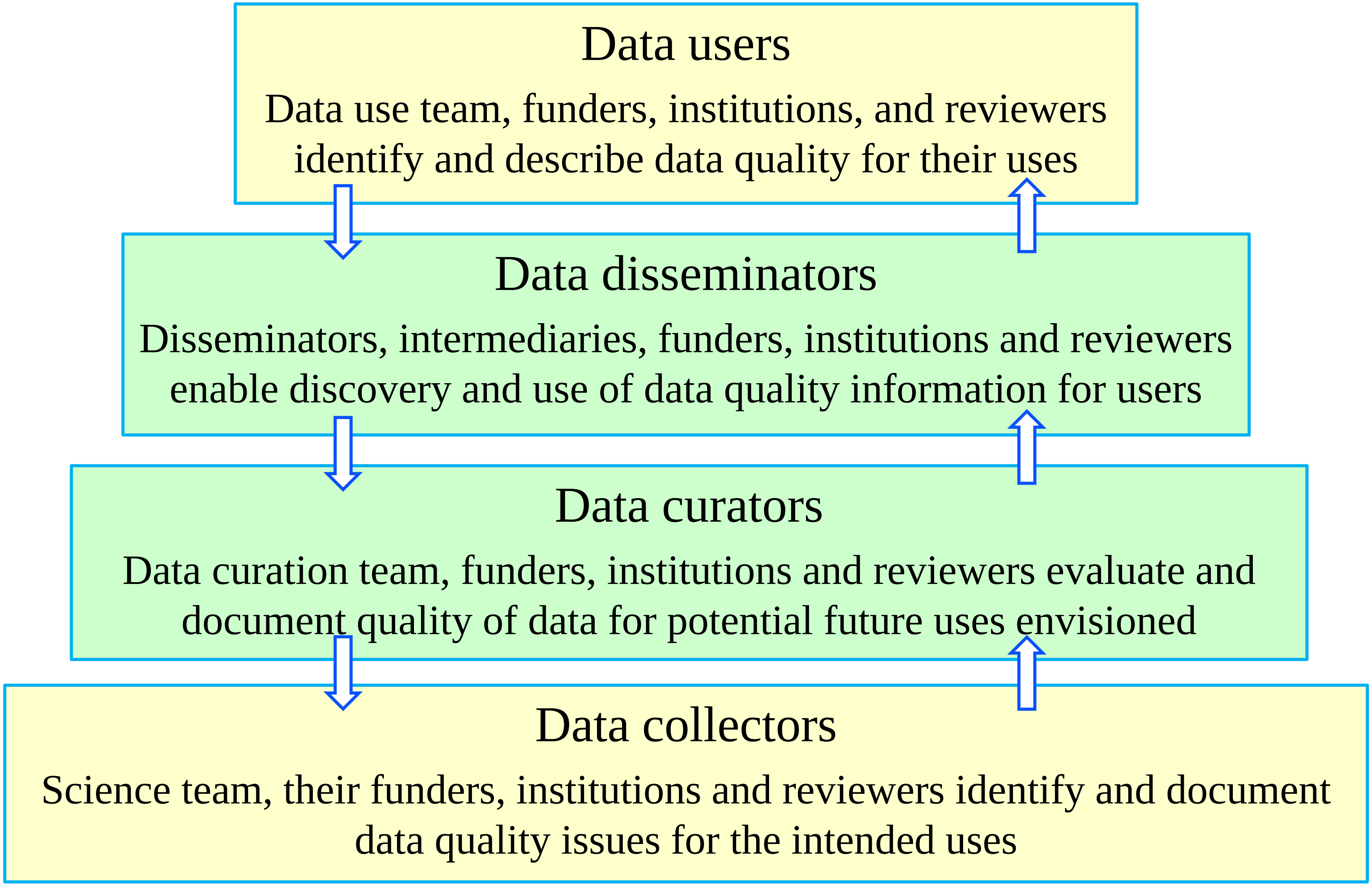
Why do we need data quality assessments?

- Identify potential opportunities for users to use the data
  - Usefulness for a particular purpose, including science studies, education, and decision-making
  - Limitations of methods, variables, or values
- Types of assessments can reflect a variety of potential uses
  - Interdisciplinary use could require multiple assessments
  - New uses of data may require new assessments
- Improve trustworthiness of the data product
  - Increase science transparency by describing quality and limitations
  - Provide independent review of data
- Describe potential for data use beyond the initial data study
  - Indicate opportunities for subsequent use
  - Identify opportunities to combine data with other data

Data Quality in Standards and Guidelines

- Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata (CSDGM) FGDC-STD-001-1998
  - Report on the data quality assessment
- ISO 19115-1:2014 Geographic information -- Metadata -- Part 1: Fundamentals
- ISO 19157:2013 Geographic information -- Data quality
  - Describing, registering, evaluating, and reporting data quality
- NASA Earth Science Data Preservation Content Specification (PCS). 2013.
  - Product quality, including methods, quality flags, uncertainty, and limitations
- ISO 14721:2012 Space data and information transfer systems -- Open archival information system (OAIS) -- Reference model.
  - Data quality reviews are reported in Representation Information
- GEO Data Management Principles Implementation Guidelines
  - Traceability, Data Quality Control, and Data Review and Reprocessing

Data Lifecycle Contributions to Data Quality



Data users

Data use team, funders, institutions, and reviewers identify and describe data quality for their uses

Data disseminators

Disseminators, intermediaries, funders, institutions and reviewers enable discovery and use of data quality information for users

Data curators

Data curation team, funders, institutions and reviewers evaluate and document quality of data for potential future uses envisioned

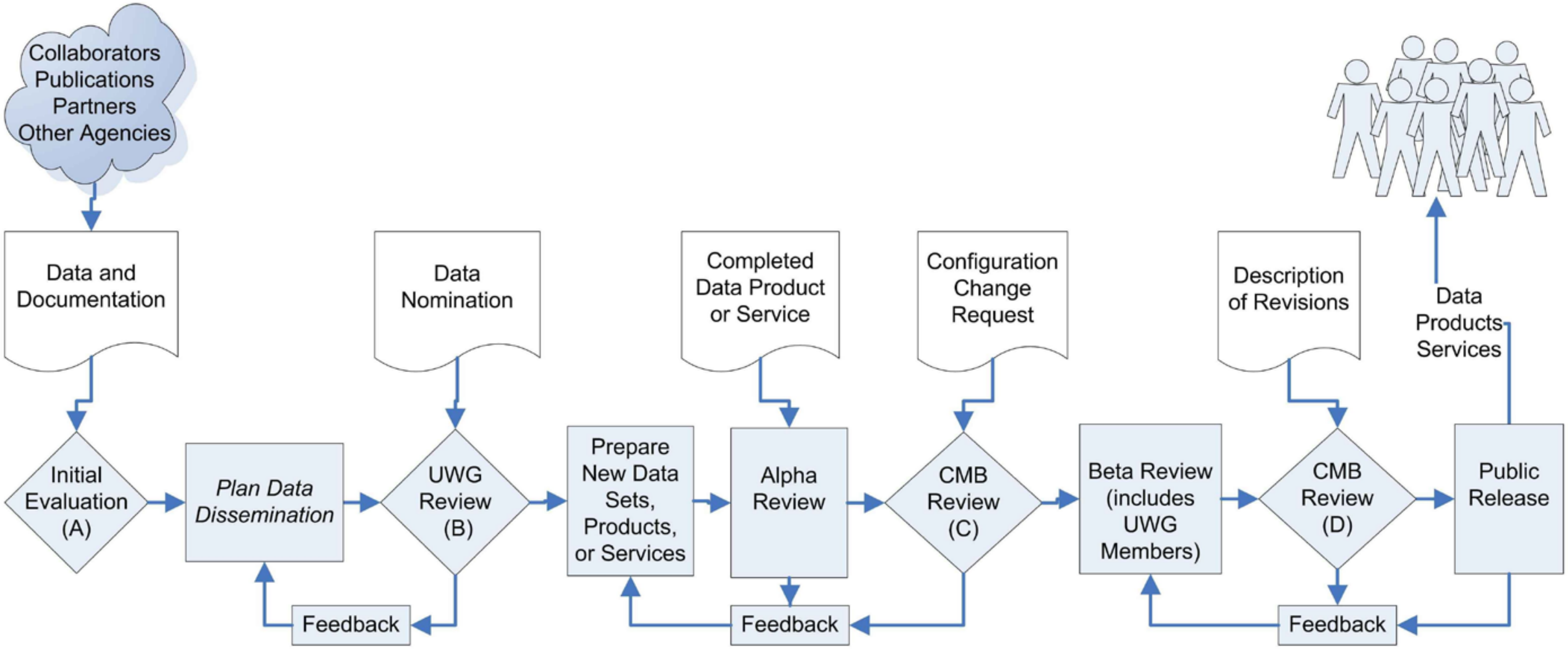
Data collectors

Science team, their funders, institutions and reviewers identify and document data quality issues for the intended uses

Data Quality in Scientific Data Study Workflow

- Conceptualize study
  - Identify and document potential uses of data
- Design study
  - Describe potential uses of data and data review and documentation process
- Collect data
  - Document data collection instruments, variables, procedures, and anomalies
- Analyze data
  - Document assumptions about data for analysis and limitations
- Deposit data
  - Submit data quality information with data
- Publish results
  - Describe limitations of data use and cite data and data quality information

Data Quality Workflow in Domain Repository



Example: NASA SEDAC data review process with feedback loops.

Source: SEDAC Guidelines for Data Review Prior to Public Release. 2013

Sample Actions for Improving the Usability of Data Quality Information throughout the Data Lifecycle

	Data Center	Science Team
Capturing Data Quality Information	**request documentation from investigators** on the extent of error introduced into data products ...	develop capabilities for investigators to **describe the extent of error** introduced …
Describing Data Quality Information	**provide** enough **publicly available information** so users do not need to contact the data center	**describe quality flags** in the data documentation and in the FAQs
Discovery of Data Quality Information	develop capabilities for users to **refine search query** results by selecting among choices of **quantifiable data quality criteria**, such as confidence levels ...	**identify quantifiable data quality criteria**, such as confidence levels and the values of quality flags, that can be used as criteria for refining search queries.
Enabling Use of Data Quality Information	provide users with a tool to **identify inputs, …, that contributed to each pixel**.	create tools to capture into a variable, **sensor inputs, … that contributed to each pixel**.

Sample Inventory of Current Implementations for Data Systems Integration

	Data Center - Investigator Communication	Metadata Creation & validation	Guidance & Instruction	Reference / Help Desk
Metadata Compliance Checker		X		
Data Quality Guide Document			X	X
Science Data Working Group	X			
Data Quality Section in Data Management Plan	X		X	
Data Management Plan / Guidelines	X		X	
FAQ Development and Analysis				X

Implications for Science Practice

- Data quality raises the stakes for data contributions
  - Effort needed to measure, document, and disseminate data quality
  - Data quality measures identify positive and negative aspects of data
- Recognition for data and data quality contributions
  - Data quality review is a scientific contribution and includes design of measures, administration of review process, conducting reviews, documenting results, and enabling use of data quality information

Logical Progression for Planning Data Quality Assessments for Specific Uses

- Plans to disseminate data should be justified by potential uses of the data

⇩

- Justification to disseminate data should include a description of the potential uses of the data

⇩

- Upon identifying a proposed use for data, plan to conduct a data quality assessment for the proposed uses of the data

Recommendations for Communicating Data Quality Information

- Each publicly accessible data set should describe its potential use
- Each claim for a potential use of data should be justified by a quality assessment
- Each data quality assessment should include a data quality indicator in data documentation or metadata
- Each data quality assessment indicator value should be defined in the data documentation or metadata
- Data users should cite data and describe their assessment of the data for the study conducted
- Update metadata and documentation to reflect reported data assessments

Taking Action to Initiate, Curate, and Disseminate Data Quality Reviews

- Archives, Repositories, and Data Centers
  - Invite user community to review data
  - Establish roles for user community members who review data
  - Include data review results when archiving data
  - Disseminate data review results with data
- Future Research Opportunities
  - Identifying "low-hanging fruit" solutions that could be reasonably executed in a relatively short time frame (current DQWG effort)
  - Identifying ways to standardize data quality practices and workflows
  - Identifying additional data quality challenges
  - Proposing new concepts to address challenges where existing solutions and best practices fall short